USING ESSAYS TO EVALUATE LEARNING AND COMPARING HUMAN SCORING OF ESSAYS TO COMPUTER SCORING SYSTEMS

by

Michelle Alissa Hudson

A thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Educational Psychology

The University of Utah

May 2016

The University of Utah Graduate School

STATEMENT OF THESIS APPROVAL

The thesis of **Michelle Alissa Hudson**

has been approved by the following supervisory committee members:

Anne E. Cook, Chair **12/08/2015**
Date Approved

Kirsten Renee Butcher, Member **12/08/2015**
Date Approved

Robert Zhiwei Zheng, Member **12/08/2015**
Date Approved

and by **Anne E. Cook**, Chair/Dean of

the Department/College/School of **Educational Psychology**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Prior research conducted by Butcher, Davies, and Cook (2015, in preparation) demonstrated that using concept maps to search within the online scientific database from the National Science Digital Library (NSDL) decreases cognitive effort over more common keyword-based searches; our purpose was to determine whether this decreased cognitive effort translated into different learning gains as measured by evaluating and scoring pre- and postessays. Teachers are one group who would benefit from more effective, less cognitively demanding ways of finding online material for their classrooms, so the participants in this study were student preservice as well as practicing inservice teachers. Using a rubric developed to evaluate the specific essays written for the Butcher et al. study, we found that participants were able to learn from online search tasks, as measured by more correct information contained in a postessay compared to a pre-essay, and a higher overall score; but this learning was not a function of which online search methods were used. The decreased cognitive effort did not lead to more learning gains as measured in this study.

Our second study compared the hand-scored results from the postessays to two computerized scoring systems: Latent Semantic Analysis (LSA) and Coh-Metrix. The purpose of such systems is to help alleviate some of the issues with scoring large numbers of essays by hand. LSA determines semantic similarity between two texts, and Coh-Metrix gives measures of cohesion within each text. LSA correlated moderately with the hand scores (0.44 for the preservice teachers and 0.38 for inservice teachers). Other

research has shown higher correlations between LSA and human graders, and because the LSA cosine scores do not show essay quality or level of correctness (only semantic similarity), they could not be substituted for the hand scores. None of the Coh-Metrix cohesion measures correlated significantly with the hand scores. This indicates that cohesion measures obtained from Coh-Metrix are not indicative of the quality of essays as determined by human scorers as given for these essays.

TABLE OF CONTENTS

I. INTRODUCTION

Using technology in educational settings is rapidly becoming the norm—in helping students learn, but also in helping teachers find and use information to prepare and supplement lessons. Most web search engines involve entering words or phrases to search for content and return what can appear to be "anything and everything" even remotely related to the topic. Educators can find it frustrating and difficult to sort through the myriad of seemingly unrelated or nonrelevant search-return material—such as advertisements, cultural references, and commercial products—to find trusted educational resources (Deniman, Sumner, Davis, Bhushan, & Fox, 2003). A way to address this problem is to restrict the search results to only those with educational or scientific content. For example, the National Science Digital Library (NSDL) is an online digital library that limits its content to high-quality educational and scientific sources (NSDL.org). NSDL was established in 2000 by the National Science Foundation (NSF) to provide digital resources for the fields of science, technology, education, and mathematics (referred to as STEM subjects) (McIlvain, 2010). The purpose of NSDL, therefore, is for users to find sources on a given topic without having to wade through extraneous and potentially unrelated material.

A user of NSDL may enter terms in a similar manner as in a Google search. However, the restricted nature of the NSDL content increases the likelihood that users (usually those with specific educational aims such as classroom teachers) will locate needed resources as compared to searches conducted with traditional, unrestricted search tools. For example, if one were to type the word "gravity" into Google, of the fourteen results on the first page, about three-fourths reference the 2013 motion picture of the same name. Of the remaining sites, perhaps only one would be of use in a classroom setting. When searching for "gravity" on the NSDL site, however, all of the returned results are about the physical (rather than the motion picture) phenomenon.

In addition to the keyword search function, NSDL includes a function that allows users to search through science Literacy Maps, which give a visualization of how different domain ideas are connected to each other (see Figure 1). Each idea is represented in a node and each node is linked together with arrows showing conceptual relationships about how the ideas build upon each other over time according to complexity and grade level. When a specific node is selected via mouse click, multiple NSDL-catalogued digital resources relating to that concept are presented (see Figure 2).

Using such maps eliminates the process of continually having to develop search terms, a process that Marchionini & White (2007) describe as effortful and limited by the amount of prior knowledge and vocabulary of the user. When using the NSDL map feature, the user simply selects an initial search term (e.g., "gravity") from a list of concept maps. If a desired term is not provided—such as a more specific aspect of the topic at hand (i.e., orbit)—a user will be led to the desired term contained within a map. In a study conducted by Hagemans, van der Meij, and de Jong (2013), one group of students was given a concept map similar to the NSDL maps to learn about a specific topic, and a second control group was asked to identify relevant concepts without the aid of a map. The learning gains of the students using the maps, as measured by a post-test, were greater than the gains for their peers in the control condition; Hagemans et al.

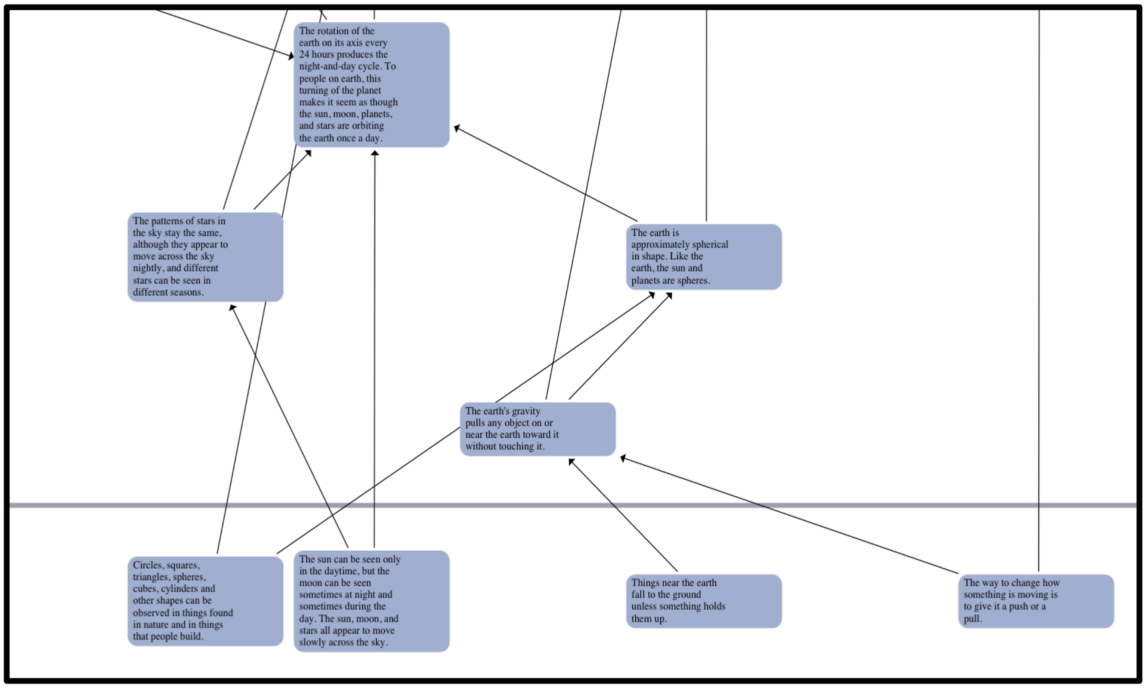

Figure 1. Part of the NSDL map for the concept of "gravity."

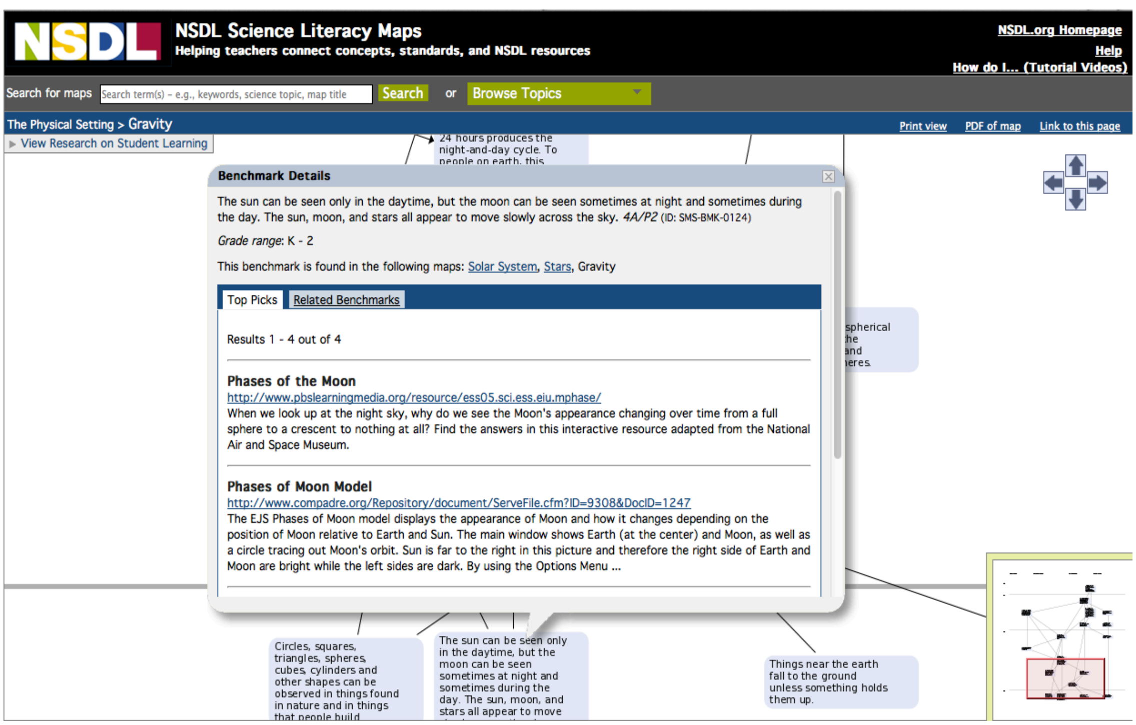

Figure 2. NSDL concept map showing the links available after clicking on a node.

concluded the use of a map that indicates an optimal route through a topic leads to improved learning. Other studies have also demonstrated that learning can be facilitated and fostered by using concept maps (Ciullo, Falcomata, Pfannenstiel, & Billingsley, 2015; von der Heidt, 2015).

Consistent with the findings just described about the benefits of concept maps, Butcher, Davies, and Cook (2015, in preparation) found that it was easier and faster for participants using the NSDL concept map format to identify and discriminate useful resources and to reject unwanted ones than it was for participants using a basic keyword search task. Butcher et al. further analyzed their results by breaking the search task into its component stages, using a framework originally proposed by Marchionini and White (2007). According to this framework, there are three basic stages involved in searching for information online: first, the formulation and reformulation of search terms; second, the examination of the result list; and third, the evaluation of the content of resources found. Using this framework, Butcher et al. examined the differences in time spent and cognitive effort exerted for each stage as a function of type of web search tool used. Forty-two participants, all of whom were students in a teacher education program, spent ten minutes using each of three search tools (NSDL keyword, NSDL maps, and Google). For each tool, the task was to try to identify four online resources that could be used in a classroom setting to supplement a lesson on the given topic. Butcher et al. used screen recordings to evaluate and calculate the time spent in each of the stages identified by Marchionini and White. The efficiency of each search was evaluated by comparing the number of accepted resources (as determined by having the participants bookmark sites they were interested in using) to the number of sites that were examined but ultimately rejected (those not bookmarked). For example, if one of the participants examined many different resources, but rejected most of them, then his or her search was classified as more inefficient that another participant who looked at only a few resources, but accepted and bookmarked most of them.

Butcher et al. discovered that participants spent more time on the formulating and reformulating phase of the task when using the NSDL map interface than when using either the NSDL keyword search or a Google keyword search. That is, they spent a significant amount of time looking at the map and deciding which nodes to select when using the maps compared to the same formulation phase of generating search terms in the keyword search conditions. Butcher et al. also found that participants spent less time looking at resources that they accepted or rejected when using maps than in the other interfaces, indicating that it took them less time to evaluate the appropriateness of the resources. This may imply that the concept maps support domain-based thinking by describing relevant domain topics and that the arrangement of these topics in an organized way facilitates quicker evaluation of the web results. Together, the findings that map-users spent more time looking at the map and nodes and less time evaluating the resource seems to indicate that domain-based thinking about the topic is facilitated by studying the maps.

Butcher et al. also found that using NSDL to constrain the returned results did not improve the keyword search task. When comparing the two keyword search conditions (i.e., Google vs. NSDL keyword), there was no difference in time spent evaluating the content, indicating that having NSDL limit results to only educational sources did not improve participants' efficiency in deciding whether a resource was of use or not. There are at least two possible reasons for this: first, participants' searches may have been specific enough not to return commercial content, or, second, the commercial content was obvious enough that the students did not need to spend time avoiding or sorting through it. In fact, the NSDL keyword search was less efficient (i.e., participants selected more resources that they ended up rejecting than accepting) than the Google search. This may be because when using Google, users are aware that not all of the results are of scientific or educational value so they are more careful about what they pick. Alternatively, there may be something inherent in the Google search returns that lead to this response. In contrast, when the results are constrained, as they are with the NSDL library, participants do not spend as much time evaluating the list of possible resources before they select a link.

In addition, Butcher et al. used eye-tracking technology to measure participants' cognitive effort, which was calculated by changes in pupil diameter, as well as number and duration of fixations made during each search stage. Butcher et al. found that in the NSDL map condition, participants expended less cognitive effort than when they had to generate their own keywords. The peak amplitude of pupil dilation (which reflects maximum amount of cognitive effort exerted) was significantly reduced for the NSDL maps condition compared to both the Google and NSDL keyword search conditions. These results support the view that the map condition resulted in searches that are more efficient.

Essay-Based Learning Assessments

Although Butcher et al. (2015) evaluated the amount of cognitive effort participants were exerting during a web search, their findings do not tell us how much participants actually learned during that same task. One common way of evaluating learning after a task like the one used by Butcher et al. is to have participants write essays before and after the task. This allows investigators to evaluate changes in participants' levels of understanding, as well as changes in misconceptions and errors in thinking. Essays are one way to evaluate the writer's level of knowledge (Foltz, Britt, & Perfetti, 1996). For example, researchers have successfully used essays to establish whether reading analogies improve learning from scientific texts (Braasch & Goldman, 2010). Participants in the Braasch and Goldman study wrote essays, and the number of correct concepts in each essay was calculated. The number of correct concepts included in the essays was related to degree of learning. Their participants also demonstrated learning by increased scores in a post-test asking questions about the target knowledge area.

Salomon, Globerson, and Guterman (1989) also used essays to measure learning by showing that computer-guided metacognition facilitates better text comprehension and writing. One month after the initial experimental sessions, participants wrote essays that were scored based on overall quality. Those who had read passages that included metacognitive questions embedded in the text had higher quality essays. The researchers concluded that the metacognitive guidance led to more internalization, which in turn facilitated better text comprehension that transferred to their writing.

Essays can also be used to demonstrate learning in web search tasks (e.g., Butcher et al., 2015). Willoughby, Anderson, Wood, Mueller, and Ross (2009) had one group of participants research a topic online for 30 minutes before writing an essay, and had separate control groups write the essays without any research. Each participant wrote about a topic for which they had a high level of previous knowledge and one where they had a low level of previous knowledge. Essays were scored based on the number of correct statements that directly answered the essay prompt. Willoughby et al. found that participants with high prior domain knowledge who were able to research the topic improved compared to participants who searched online on a low-knowledge topic and those in the control groups who did not research the topic (both high and low previous knowledge). Searching the Internet in a low-knowledge domain did not lead to better performance than control groups who did not search. This shows that an important part of learning from online material depends on the learner's prior knowledge base as well as the search itself.

One advantage of using essays over other methods, such as multiple-choice-type evaluations, is that essays can expose writers' misconceptions and errors in thinking. Although it was not included as part of the Butcher et al. (2015, in preparation) analysis, the participants in their study also wrote a pre- and postessay about each topic researched. By evaluating the essays, it is possible to determine not only the amount of knowledge gleaned from the task, but whether or not a specific search interface facilitates learning and understanding of a given topic. For example, based on the findings of Butcher et al., we know that participants expended less cognitive effort when using the NSDL maps than when they conducted keyword searches. There may be a negative relationship between the cognitive effort needed to find online resources and amount of learning. That is, the decreased cognitive load for the maps may lead to increased learning, and, if so, this difference in learning outcomes may be evident in the participant essays. Thus, as more cognitive effort is needed for a search task, this may result in reduced capacity left for processing and learning the material.

Scoring Essay-Based Assessments

Hand-Scoring

The "gold standard" method of evaluating and critiquing essay compositions is expert human reading (Landauer & Psotka, 2000). However, there are several problems associated with human scoring. First, essay scoring is labor-intensive and can become too expensive when large numbers of essays need to be evaluated (such as for standardized testing). Second, it can also be difficult to analyze essay content. Determination of essay quality is often based on the degree of match between what the grader believes to be important in the domain of interest and what was written in the essay. What information is deemed important is determined by the grader, who is often unable to account for all of the source material used by the writer (Foltz, Britt, & Perfetti, 1996). Third, potentially irrelevant aspects of an essay (such as grammar) may also influence an essay's score. Townsend et al. (1993) examined superficial (i.e., not content-based) aspects and found that just changing the introduction of an essay but not any of the content improved an essay's overall holistic score (measured on a lettered scale) as well as additional scores (rated from "poor" to "excellent") on six other characteristics of the essay, such as organization and clarity. Fourth, there is often low reliability between human scorers (Attali, Lewis, & Steier, 2012). In a large review comparing multiple reliability studies, it was found that the mean reliability estimates for essays rated by two scorers on a holistic measure was 71% (Breland, Bridgeman, & Fowles, 1999). To eliminate these types of errors and to make the process of essay evaluation more efficient, there is increasing interest in using automated scoring methods such as Latent Semantic Analysis (lsa.colorado.edu) and Coh-Metrix (cohmetrix.memphis.edu).

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a computer programs that evaluate essays, and it has been reported to be as good as human scorers (Graesser, Li, & Feng, 2013). LSA is designed to compare large bodies of text to each other to determine their semantic similarities, and it can do so without relying on exact word overlap. LSA creates a very high dimensional space—which can include hundreds of dimensions—from a corpus of interest formed from several texts in a given domain. The LSA matrix is constructed where columns are words and rows are documents. Matrix cells are the frequency of each word within each document (Dumais, 2003; Kintsch, 2001). LSA looks at word usage; things such as word order and syntax are not taken into account. This matrix is transformed into a high-dimensional space. LSA claims that word meanings can be represented as vectors in this space (Kintsch, 2001). Newly inputted texts are then compared to this dimensional space and matched based on semantic similarity and not on exact word overlap or matching. For example, the words "teacher" and "educator" are highly related to each other and co-occur in many contexts. However, we can imagine a situation where one document contains just the term "teacher," and another document contains just "educator," but the two terms never both appear in the same document. LSA would still consider these documents to be semantically related because of other co-occurring words (e.g., school, education, students, classroom, etc.) that frequently appear with both of these terms in LSA's database (Dumais, 2003). LSA does not use constructed resources like dictionaries to determine semantic relationships; it calculates its comparisons by using only large bodies of text that are assembled by researchers to create a domain base for a specific topic (Dumais, 2003). Therefore, when a written essay is entered into LSA, its words and sentences are compared to this larger “training corpus,” or the large body of text within the LSA database, and an analysis on the words’ relations to each other is calculated (Foltz, Kintsch, & Landauer, 1998). To use LSA to evaluate essays, a user can identify an area of interest within the semantic space by supplying a model essay or set of comparison documents in the correct domain; LSA compares the to-be-graded essays against these documents.

Foltz (1996) found that LSA was as reliable in judging quality of essays as were raters scoring the same essays by hand. In his study, four graders, who were familiar with the subject content, evaluated and graded essays based on which sources were cited and used in each essay, and on the quality of the information cited. Essay “quality” in this study referred to the degree of semantic similarity between the essay and the texts on which it was based. LSA was just as reliable as the expert graders in characterizing the quality of the essays, meaning that LSA was as highly correlated with each of the human raters as they were with each other.

Although LSA seems to be a reliable alternative to human scoring, one possible confounding issue that may arise when using LSA to evaluate essays is word count. Layfield (2012) found that longer essays have more accurate semantic similarity comparisons than shorter essays. This is because, as the number of words increase, there is more semantic information available and therefore more relationships between words to consider. In Layfield’s research, two groups of students were given the same question to answer, but one group was given half a sheet to fill and the other group was given a full page. Most students tried to fill the space given to answer the question because the essay was part of their final exam; consequently, those given more space wrote longer essays. It was found that LSA rated the longer essays as more semantically similar to the LSA model essays contained within the LSA database. These essays also had superior performance overall, most likely due to increased elaboration of concepts allowed by the additional space. Similarly, Rehder et al. (1998) argued that essay length, at least for essays with more than sixty words, is correlated with the amount of knowledge the participant has about the subject in question. Essays with fewer than 60 words were not predictive of knowledge level. As essay length increased up to 200 words, the LSA cosine became increasingly predictive but with decreasing marginal returns, meaning that accuracy in determining knowledge level in essays with more than 200 words may be negligible. Rehder et al. also found a very low correlation between word count and LSA vector lengths, which can be thought of as how much LSA knows about a given topic, or "position within an n-dimensional space" (Rehder et al., 1998, p. 341). Cosines are derived from the angle between the vectors of each text. Longer essays may not be correlated with vector lengths because of the increases in nonessential and other filler words rather than increases in essential, topic-related words.

Coh-Metrix

A second computerized essay-scoring program is Coh-Metrix, which analyzes texts on over 200 measures. It was developed as a tool that will understand natural language (Graesser, McNamara, Louwerse, & Cai, 2004). A large part of Coh-Metrix examines two things: cohesion and coherence, which are different types of "connectedness" among text elements in discourse (Baig, 2012). *Cohesion* includes physical properties of the text (e.g., grammatical and lexical) that facilitate understanding. Cohesion can be broken down into "referential cohesion" and "causal cohesion." Referential cohesion is the degree of word or concept overlap across sentences, paragraphs, or the entire text. Causal cohesion is the degree to which causal relationships are explicitly drawn in the text, usually by connectives (i.e., *because, so, therefore*, etc.) (McNamara, Crossley, & McCarthy, 2010). It has been found that readers with low prior knowledge of a subject can comprehend highly cohesive texts (those with many textual cues) more easily, whereas readers with more previous knowledge are able to glean more from texts when there are cohesion gaps in the text that require them to make inferences and connections themselves (Graesser et al., 2004). When readers are required to make inferences from the text, the more connections between the text and prior knowledge the reader can make leads to more reasoned mental representations. If these connections cannot be made based on prior knowledge, more cohesion cues in the text are needed (McNamara, 2001).

Coherence refers to the general organization of the text that leads to these mental representations. A text is coherent when a clear mental representation can be formed. This representation depends on the previous knowledge and experience the reader brings to the text (Graesser et al., 2004). "Coherence is the semantic unity that flows throughout the text and makes it an overall 'meaningful whole'" (Baig, 2012, p.100). Thus, cohesion is a textual construct, whereas coherence is a psychological one (Graesser, McNamara, & Louwerse, 2003).

Although cohesion and coherence represent different constructs, they are highly correlated, and cohesion can help to facilitate the development of a coherent mental representation of the text by providing context cues to help the reader (McNamara, Crossley, & McCarthy, 2010). A text's cohesion can either help or hurt the coherence of the text. “Markers” in the text (or explicit words, phrases, or other cohesive features that guide the reader) help the reader make coherent connections with the rest of the text (Graesser, McNamara, & Louwerse, 2003). The addition or deletion of cohesion cues can help or hurt the text’s coherence for a reader. The idea that text comprehension depends in part on a text’s cohesion led to the development of the program Coh-Metrix, which analyzes text on 50 different types of cohesion relations, as well as over 200 additional measures of language, text and readability (108 of which are available online) (Graesser et al., 2004; McNamara, Ozuru, Grasser, & Louwerse, 2006). Coh-Metrix was shown to correctly differentiate between texts of different cohesion levels that had been intentionally altered to have low coherence by presenting sentences in alternative orders and by interrupting the temporal flow of the story (McNamara et al., 2006). In this study, the more traditional measures of text difficulty (such as Flesch-Kincaid Grade level) incorrectly labeled the high-cohesive text as more difficult.

Coh-Metrix also reports LSA cosines, but it does not require a comparison essay or set of text supplied by the user because Coh-Metrix is based on within-text measures only. Each essay is compared against itself by semantic overlap between sentences and paragraphs rather than semantic similarities to a second text.

Crossley and McNamara (2010) found that for human raters, the absence of cohesive devices (word overlap, connective words, etc.) was associated with a more coherent mental representation of the text. This is the opposite of what Coh-Metrix assumes to be true. Crossley and McNamara used expert raters in their study, however. Thus, it is possible that raters’ background knowledge of the subject influences their assessments of essay quality more than cohesive devices. This refers back to the finding that high-knowledge readers do better with low-cohesive texts because they are able to create their own coherent understanding of the text (Graesser et al., 2004). Likewise, McNamara, Crossley, and McCarthy (2010) found that there was no evidence that higher scored essays were more coherent. Thus, a major theoretical basis for Coh-Metrix—coherence—may not actually be considered in human grading when evaluating essay quality.

Research Questions

Given that essays are a valid assessment of student learning (Landauer & Psotka, 2000), it should be possible to use essays to determine whether web-search task conditions result in differences in learning. Butcher et al. (2015) found that using NSDL maps instead of a key word search reduced cognitive effort. However, it is not clear from the results they reported whether those reductions in cognitive effort are associated with differential learning gains. It may be possible to address this issue by examining the essays written by the participants of the Butcher et al. study for each of the three search conditions. This is the overall goal of this research.

For the first part of this research, human graders analyzed the pre- and postessays written by the participants in the Butcher et al. (2015) study in each of the three search conditions (NSDL map, NSDL keyword search, and Google keyword). Each essay was broken down into idea units that were each evaluated as either being a correct, incorrect, or an extraneous statement. Error revision scores were also calculated by evaluating how mistakes in the pre-essay were fixed in the postessay. In addition to these measures, an overall score was applied to all pre- and postessays. The overall score was an indication of how much of the essay contained correct and relevant information and how thoroughly the prompt was discussed. It was also a gauge of the essay's overall organization and flow. (See Appendix A for the complete rubric used to evaluate the essays). Inter-rater reliability scores were calculated to make sure the rubric was reliable for all users, since human scorers are not always reliable (Attali, Lewis, & Steier, 2012). This allowed for the evaluation of the following research questions:

1. Are the scores given for all measures consistent across multiple scorers?
2. Does the type of search task used influence learning, as measured by increases in number of correct statements from pre- to postessays?
3. Does the type of search task used influence learning, as measured by decreases in the number of incorrect and extraneous statements from pre- to postessays?
4. Does the type of search task influence a participant's success in revising incorrect information, as measured by the extent to which errors in the pre-essay are "fixed" in the postessay?
5. Does type of search task used influence learning, as measured by changes in the overall score from pre-essay to postessay?

Because learning in any context depends on how much background knowledge one has, it may be that preservice and practicing teachers learn differently as a function of search task. These two groups will be compared to answer the question:

6. Are learning gains, as measured by the differences between pre- and postessays on the measures in the above questions, comparable for both preservice and inservice teachers?

The second study focuses on comparing human scores to those generated by computerized systems. Therefore, we compared the overall score produced by hand scoring of the essays to the scores produced by LSA and Coh-Metrix. Thus, the last research questions are:

7. Is the overall score given by hand scoring in the postessay correlated with LSA cosines?
8. Is the overall score given by hand scoring in the postessay correlated with measures of Coh-Metrix that evaluate cohesion?

II. STUDY ONE: MEASURING LEARNING GAINS

The first goal of this research was to determine if lower cognitive effort in the NSDL map condition, as found by Butcher et al. (2015), was associated with increased learning, relative to the Google and NSDL keyword conditions. Lower cognitive effort used in searching for material may indicate that there are more cognitive resources available to learn from the material found. In general, pre- to postessay learning was demonstrated by, first, an increased number of correct statements; second, a decreased number of incorrect statements and "other" (or off-topic, extraneous, etc.) statements; third, effective error revision scores; and, finally, a higher overall score in the postessay. Whether or not experience level—referring to whether a participant is a preservice or practicing teacher—influences pre- versus postessay results was also assessed. It was expected that practicing teachers would begin with higher pre-essay overall scores than preservice teachers, and therefore would not experience as dramatic learning gains as the preservice teachers, regardless of the search tool used. Inter-rater reliability in the scoring of all measures was evaluated by calculating Intraclass Correlation Coefficients.

<u>Method</u>

Participants

The participants were the same preservice teachers who participated in the Butcher et al. (2015) study. They consisted of 42 students in the preservice teacher education program at the University of Utah (6 males, 28 females, mean age = 24). Most of the participants were near to completing their undergraduate teaching degrees, with 70% in the third or fourth year of their studies at the university. All participants were compensated $50 for the three-hour study session.

In addition, 18 in-service teachers completed the same tasks as reported in Butcher et al. (2015). Ten of these teachers had a master's degree and the other eight held a bachelor's. All taught science (e.g., Biology, Chemistry, and/or Earth Science) in either middle school (eight of the teachers) or high school (ten of the teachers). Three had only taught for 1-2 years, six had taught for 3-4 years, three had taught for 5-10 years, and the remaining eight teachers had taught for 10 or more years.

Materials

The materials used for the web-search tasks are the same as used by Butcher et al. (2015), and are described in the following sections.

Demographic Survey

A 33-item demographic survey was given to each participant to assess their self-reported experiences with computers, weekly time online, comfort with using the web, and perceived success of finding desired information during online searches. Previous researchers have found that experience with computers and computer searches are significant predictors of individuals' perceptions of search engines (Laiw & Huang, 2003).

Instructional Task

Participants were given ten minutes to find up to four digital resources to supplement instruction in a classroom setting related to concepts in the state core curriculum. The three topics used were plate tectonics, cells, and the water cycle. Pilot testing found that preservice teachers in particular needed specific information about classroom context, the nature of the students, and instructional goals to make the best decisions about what digital resources to select. Each task, therefore, had three parts: a relevant instructional standard and objective from the state core curriculum, specific information about the classroom context, and a clear goal for the online search. Each of the three search tasks was assigned to one of the following classroom contexts: 1) Find digital materials on the water cycle to support struggling learners; 2) Find digital, interactive material on cell biology to engage all learners, including students who are English language learners; and 3) Find digital materials to help all students visualize how plate interactions relate to natural phenomena. (Please see Appendix B for an example of the instructions for performing a search task.)

Web-Search Tools

All participants used three search tools, one for each topic. First, Google.com was used as a comparison search engine due to its wide use and familiarity. All participants in the Butcher et al. (2015) study reported using Google frequently for web searches. The other search tools were both part of the National Science Digital Library (NSDL) database, which collects and catalogues online educational resources in the fields of science, technology, engineering, and mathematics. The second tool was NSDL keyword search, which operates similarly to a Google search but with results constrained to those of educational relevance (as compared to more varied content available on the unrestrained Web). The third search-tool was the NSDL concept maps, which offer a graphical search interface in the form of a node-link diagram. Clicking on specific nodes brings up relevant resources that are displayed as a hyperlinked resource with a title, URL, a short description of page content, and a list of relevant keywords. Nodes are connected together via lines to show the relationships between concepts and the gradual expansion of the topic over age-level and increasingly complex concepts. Users do not continually need to use keywords to find new content, but can simply exit out of a given result list to choose a different node on the map.

Procedure

An essay prompt was given for every participant to respond to before and after the web-search task (please see Appendix C for the prompts). They were instructed to write what they knew about one of the three science topics (plate tectonics, cells, or the water cycle). Each participant was given five minutes to write his or her response. They each then had ten minutes to complete a web search. They were asked to bookmark resources they deemed useful for instruction in a classroom and that they would choose to use. This was followed by a 10-minute learning task. Participants were asked to go back through the resources they bookmarked to read through the material with the purpose of learning from them. They were also reminded of the essay prompts at this time (please see Appendix B). They were then given five additional minutes to respond again to the same essay prompt they saw at the beginning of the activity. This second essay was written from scratch and the participants were not shown their original response. This was repeated for the remaining search tools and topics. Each topic was randomly assigned to one of the three search methods, and the design was counterbalanced across topics and search conditions.

Essay Scoring Rubric

Each essay was broken down into idea units. Each independent clause or sentence was counted as an idea unit. Statements that included lists of items were also broken down into separate idea units for each item (i.e., the sentence, "The parts of the water cycle are precipitation, evaporation, and condensation" would include three idea units.) To evaluate each essay by hand, a scoring rubric was developed based on the specific needs of the Butcher et al. (2015) study essays. A list of common facts for each of the three topics was included as a reference for the scorers, but was not considered inclusive. (See Appendix A for the rubric used). Each separate idea unit was evaluated as correct, incorrect, or "other." An idea unit was marked as "other" if it was off-topic or if it was too vague to be marked either incorrect or correct.

Error revision scores were also calculated. If an error was present in the pre-essay, the postessay was evaluated to determine how the error was addressed. There were three possibilities: first, the error could have been a *fixed error*, or an error that was corrected in the postessay; second, the error could have been a *same error*, indicating that the same error in the pre-essay is still present in the postessay; or, third, it could be a *not-addressed error*, indicating a situation where a pre-essay error is neither corrected nor still present but instead the issue is absent from the postessay altogether. In addition, postessays could contain *new errors*, or errors unique to the postessay.

An overall score (1-5) was also given to each essay. This overall score was designed to evaluate the essay's level of correctness as a whole, while also considering readability and fluidity. For example, an overall score of 1 indicates that there is little if any relevant domain content and may contain many errors; an overall score of 3 indicates that much of the information is correct, but it may be lacking in detail or proper explanation; and finally, an overall score of 5 indicates that the topic is well covered and errors, if any, were very minor.

Results

Interrater Reliability

A second rater scored 20% of the essays to determine the reliability of the scoring rubric. Intraclass correlation coefficients (ICC) were determined for all measures. The ICC for overall score was significant (r=0.89, p<0.05). ICC for correct statements was also significantly correlated (r=0.81, p<0.05), as were ICC for "other" statements (r=0 .90, p<0.05). Significant intraclass correlations were also found for incorrect statements (r=0.90, p<0.05). Reliability was also calculated for 20% of the essays that contained errors, and interrater reliability was verified for each of the corrections of errors in the postessays. Each was statistically significant (ICC for "Fixed errors": r=0.84, p<0.05; ICC for "same errors": r=0.88, p<0.05; ICC for "new errors": r=1.00, p<0.05; and ICC for "not addressed errors": r=0.93, p<0.05).

Idea Unit Results

Preservice and inservice teacher essay scores were analyzed separately. We compared the two essays (pre and post) to the three search tools (Google, Keyword, and Maps) separately for correct, incorrect, and "other" statements, as well as the overall score. Therefore, each scored value was assessed with within-subjects 2 X 3 ANOVA (pre- and postessay X the three search-tools used). In order to account for possible variance due to the order in which participants used the three search tools, a counterbalancing group was included as a between-subjects variable. Descriptive statistics for all conditions are reported in Table 1.

Significantly more correct statements were made in the postessays than in the pre-essays; this was true for both the preservice teachers ($F(1,34)=67.01$, $p<0.05$, $\eta^2=0.66$) and inservice teachers ($F(1,12) = 10.57$, $p<0.05$, $\eta^2=0.47$). However, there was no main effect of search tool (Maps, Keyword, or Google) (preservice: $F(2,68)=0.35$, $p=0.70$; inservice: $F(2,24)=1.79$, $p=0.19$). There were also no significant interactions between test time and search tool used (preservice: $F(2,68)=2.22$, $p=0.12$; inservice: $F(2,24)=0.39$, $p=0.68$).

There was no main effect of test time for incorrect statements (preservice: $F(1,34)=0.20$, $p=0.57$; inservice: $F(1,12)=0.31$, $p=0.59$). There was also no main effect of search tool (preservice: $F(2,68)=1.13$, $p=0.03$; inservice: $F(2,24)=0.27$, $p=0.78$). The interaction was also not significant (preservice: $F(2,68)=1.51$, $p=0.23$; inservice: $F(2,24)=0.32$, $p=0.73$).

Likewise, for "other" statements there was no main effect of test time (preservice: $F(1,34)=0.52$, $p=0.46$; inservice: $F(1,12)=0.00$, $p=1.00$) or for search tool (preservice: $F(2,68)=2.37$, $p=0.10$; inservice: $F(2,24)=0.51$, $p=0.61$). Once again, there were also no significant interactions between test time and search tool (preservice: $F(2,68)=1.26$, $p=0.29$; inservice: $F(2,24)=0.09$, $p=0.91$).

Table 1

Mean and standard deviations for preservice teachers and inservice teachers for types of idea units and overall scores.

		Preservice Teachers		Inservice Teachers	
Type of Score	**Search Tool**	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
Pre-essay Correct Idea Units	NSDL Maps	7.98	3.97	15.67	5.89
	NSDL Keyword	6.61	3.86	14.67	6.66
	Google	7.00	4.52	15.00	4.34
Postessay Correct Idea Units	NSDL Maps	10.38	3.56	17.61	5.08
	NSDL Keyword	11.24	5.20	16.22	5.55
	Google	10.93	4.80	15.61	3.87
Pre-essay Incorrect Idea Units	NSDL Maps	0.74	1.01	0.44	0.98
	NSDL Keyword	1.10	1.39	0.44	0.62
	Google	1.38	1.74	0.28	0.46
Postessay Incorrect Idea Units	NSDL Maps	1.08	1.70	0.33	0.77
	NSDL Keyword	0.83	0.95	0.56	0.98
	Google	1.10	1.28	0.50	0.79
Pre-essay "Other" Idea Units	NSDL Maps	2.24	2.43	1.00	1.57
	NSDL Keyword	2.51	2.51	1.33	1.71
	Google	1.71	1.80	1.28	1.27
Postessay "Other" Idea Units	NSDL Maps	2.90	3.36	1.00	1.19
	NSDL Keyword	2.29	1.94	1.44	1.76
	Google	2.07	1.52	1.17	1.10
Pre-essay Overall Score	NSDL Maps	2.17	0.66	3.28	0.75
	NSDL Keyword	2.24	0.66	3.22	0.73
	Google	2.33	0.79	3.44	0.78
Postessay Overall Score	NSDL Maps	2.75	0.67	3.61	0.70
	NSDL Keyword	2.71	0.60	3.39	0.70
	Google	2.85	0.61	3.72	0.89

Error Revision Scores

Because error revision scores are only evaluated in the postessays, we only tested for an effect of search tool. There was no main effect for search tool on "fixed errors" (preservice: $F(2,68)=1.05$, $p=0.35$; inservice: $F(2,24)=1.00$, $p=0.38$), "same errors" (preservice: $F(2,68)=0.35$, $p=0.70$; inservice: $F(2,24)=0.90$, $p=0.42$), or "not-addressed errors" (preservice: $F(2,68)=1.91$, $p=0.16$; inservice: $F(2,24)=0.70$, $p=0.51$). There was also no main effect for search tool for "new errors" (preservice: $F(2,68)=0.81$, $p=0.45$; inservice: $F(2,24)=0.59$, $p=0.57$).

One possible reason for the nonsignificant effects for error revisions may have been that some essays did not contain any errors. Of the 60 total participants, only two did not make any errors in any of their six essays. However, there were 11 participants who made errors in only one topic (and made zero errors in the other two), 28 who made errors in two topics, and 19 who made errors in all three topics. Overall, inservice teachers made far fewer errors then the preservice teachers (51% of all inservice essays contained zero errors, whereas only 22% of preservice essays contained zero errors). In order to control for the possibility of essays without any errors to interfere with the results for error revision scores, the analyses were rerun as a one-way ANOVA with only preservice teachers who made an error in the given topic. Each of the three essay topics (plate tectonics, cells, and the water cycle) was examined independently to determine if there were significant differences in error correction as a function of search task. Participants who did not make an error about that topic were eliminated. These results were the same as in the larger analysis—there were no significant effects of search task on error revision (see Table 2).

Table 2

One-way ANOVA results for postessays containing at least one error or error revision.

Essay Topic	Error Category	F	P
Plate Tectonics	Fixed Errors	$F(2,36) = 1.03$	$p>0.05$
	Same Errors	$F(2,36) = 1.02$	$p>0.05$
	Not Addressed Errors	$F(2,36) = 0.15$	$p>0.05$
	New Errors	$F(2,36) = 0.88$	$p>0.05$
Cells	Fixed Errors	$F(2,31) = 0.47$	$p>0.05$
	Same Errors	$F(2,31) = 0.23$	$p>0.05$
	Not Addressed Errors	$F(2,31) = 2.89$	$p>0.05$
	New Errors	$F(2,31) = 1.15$	$p>0.05$
Water Cycle	Fixed Errors	$F(2,22) = 3.27$	$p>0.05$
	Same Errors	$F(2,22) = 3.17$	$p>0.05$
	Not Addressed Errors	$F(2,22) = 1.37$	$p>0.05$
	New Errors	$F(2,22) = 3.06$	$p>0.05$

Overall Scores

There was a significant main effect of test time for overall scores; participants had higher scores on postessays than on pre-essays (preservice: $F(1,34)=33.03$, $p<0.05$, $\eta^2=0.49$; inservice: $F(1,12)=12.25$, $p<0.05$, $\eta^2=0.51$). There was no significant main effect of search task on the overall score (preservice: $F(2,68)=1.08$, $p=0.35$; inservice: $F(2,24)=1.10$, $p=0.35$). There were also no significant interactions between test time and search task for the overall scores (preservice: $F(2,68)=0.60$, $p=0.55$; inservice: $F(2,24)=0.40$, $p=0.68$).

Group Comparisons

T-tests were used to compare the difference between preservice and inservice teachers on all of the scored measures. (Please see Table 1 for descriptive statistics on measures reported as a function of teacher group.) On average, preservice teachers included 7 correct statements in their pre-essays and inservice teachers included 15. This difference was significant for each of the search tasks (NSDL Maps: $t(58)=-5.92$, $p<0.05$, $d=1.70$; NSDL keywords: $t(57)=-5.86$, $p<0.05$, $d=1.70$; Google: $t(58)=-6.36$, $p<0.05$, $d=1.82$). There was also a significant difference in correct statements for the postessay—preservice teachers averaged 11 statements, inservice teachers 16—across the three search tasks (NSDL Maps: $t(56)=-6.25$, $p<0.05$, $d=1.80$; NSDL Keyword: $t(57)=-3.32$, $p<0.05$, $d=0.96$; Google: $t(57)=-3.65$, $p<0.05$, $d=1.05$). For incorrect statements, preservice teachers averaged 1 error in both the pre- and postessay and inservice teachers averaged 0.4 errors in the pre-essay and 0.5 errors in the postessay). These differences were only significant at the $p<0.05$ level for pre-essays in the Google condition ($t(58)=2.64$, $p<0.05$, $d=0.76$). For "other" statements, preservice teachers averaged 2 other statements in the pre-essay and 2.5 in the postessay; inservice teachers averaged 1 "other" statement in both essays. These differences were only statistically significant for the postessay statements in the NSDL Maps and Google conditions (NSDL Maps: $t(56)=2.32$, $p<0.05$, $d=0.67$; NSDL Keyword: $t(57)=1.60$, $p=0.12$; Google: $t(57)=2.27$, $p<0.05$, $d=0.65$); all other contrasts yielded $p\geq0.05$. Finally, for the overall scores, preservice teachers had an average score of 2 in the pre-essay and 2.8 in the postessay, whereas inservice teachers had an average of 3.3 in the pre-essay and 3.5 in the postessay. The differences between the two teacher groups were significant for both the pre-essay tasks (NSDL Maps: $t(58)=-5.73$, $p<0.05$, $d= 1.64$; NSDL Keyword: $t(57)= -5.06$, $p<0.05$, $d=1.46$; Google: $t(58)=-5.02$, $p<0.05$, $d=1.44$), and the postessay tasks (NSDL Maps: $t(56)=-4.47$, $p<0.05$, $d=1.29$; NSDL Keyword: $t(57)=-3.81$, $p<0.05$, $d= 1.10$; Google; $t(57)=-4.33$, $p<0.05$, $d=1.25$).

Discussion of Study One

Butcher et al. (2015) found that different methods of finding online information led to differences in time spent and cognitive effort exerted. Specifically, when using the NSDL Concept Map interface, users were able to identify useful websites more quickly and used less cognitive effort. The purpose of this study was to determine whether this cognitive efficiency with the map condition was associated with increases in learning gains as evidenced by increases in correct statements and decreases in incorrect statements in written essays. Lower cognitive effort used to find resources might indicate that more of those same resources can be used to learn from the material found. It is also a possibility that the amount of cognitive effort used in finding resources is not related to how much learning took place as measured by an essay. Based on the results of this study, there was no evidence of differences in learning gains as a function of search task for any of the measures analyzed.

However, participants did appear to learn regardless of search task condition; they showed a significant increase in the number of correct statements and in their overall score from pre-essay to postessay. Nevertheless, there was not a significant change in incorrect or "other" statements, or in the revision of the incorrect statements from pre-essay to postessay. This may be due to the short nature of the essays. The average length of the postessays was 138 words for preservice teachers and 161 words for inservice teachers. With such little material being written, it may be the case that participants were focusing on what knowledge they felt certain of instead of areas of the topic with which they were less familiar. These results could also be a factor of time, as participants were only given five minutes to write each essay. Perhaps if participants were given more time to write longer essays, it would be possible to see changes in knowledge that would show up in error revisions. We might also see an effect of search tools if participants were given more time to use each search tool.

T-tests comparing the preservice teachers to the inservice teachers showed that there were significant differences between the groups in terms of correct statements and overall scores. The inservice teachers had more correct information and higher scoring essays, even in the pre-essay. This was expected due to their higher amount of experience and previous knowledge about the science topics in question. The short nature of the essays and the limited time given to write them might explain the lack of a difference in incorrect statements, because as mentioned previously, most essays for each group contained only one error, if any at all. It is possible that inservice teachers may have experienced a ceiling effect based on how much information they are able to provide within a 5-minute time limit. If participants were able to write all they knew about a subject without a time limit, for both pre- and postessays, it may have been possible for these teachers to express any new information they studied in addition to the possibility of just rewriting the same (correct) information contained in the pre-essay.

III. STUDY TWO: COMPARING HAND-SCORES TO COMPUTERIZED SCORING SYSTEMS

Scoring essays by hand can be a tedious business! Therefore, it is of interest to determine whether automating the process by using a computerized tool can replicate how a teacher or other rater would score a given essay. To this end, several programs, including Latent Semantic Analysis (LSA) and Coh-Metrix, have been developed. The purpose of Study 2 was to determine whether the relationship between the overall score given in Study 1 is similar to scores assigned by LSA and Coh-Metrix. For this study, we are not interested in the effect that search task had on the essays, so scores were collapsed across search task to determine how each essay's overall score correlates with these computerized scoring systems.

As discussed previously, LSA compares each essay to a standardized essay to determine semantic similarity (Dumais, 2003; Foltz, Kintsch, & Landauer, 1998). If human-derived scores correlate strongly with LSA cosines, LSA could be used to differentiate between high- and low-scoring essays with much less time and effort.

One of the main tasks for Coh-Metrix is to determine the "connectedness" of each text by how cohesive and coherent it is (Baig, 2012). We are interested in determining whether there are correlations between Coh-Metrix cohesion scores with the overall hand scores. Because cohesion and coherence are partially dependent on the knowledge level of the reader (McNamara, 2001), we are interested in the differences in performance between preservice teachers and inservice teachers and if their knowledge differences will be reflected in the scores given by LSA and Coh-Metrix.

Method

Participants

The participants were the same preservice and inservice teachers from Study 1.

Materials

Only the postessays from Study One were used in this study.

Hand-Graded Essays

The overall score given to each postessay was used as a comparison for the computerized scores based on the rubric guide. These are the same scores as used in Study 1.

LSA

Essays entered into LSA were compared to a main text, which was written as a "perfect" response to the prompt and included as much information about the given topic as possible (please see Appendix D for the comparison essay for the topic of cells). Each participant essay was entered into the LSA website (lsa.colorado.edu) under the "one-to-many" analysis. LSA then evaluated each essay by giving a cosine score indicating the degree of similarity to the main text by using LSA's "document-to-document" comparison option. For each topic, the LSA topic space "General Reading up to 1st Year College (300 factors)" was used.

Coh-Metrix

Each essay was also analyzed by the Coh-Metrix website (tool.cohmetrix.com). Coh-Metrix gives an abundance of results (108 different measures in all). Of these, only a subset of selected subcategories of cohesion were used:

- **Referential cohesion**: This measures the degree to which words and ideas overlap across sentences and the text as a whole. Low referential cohesive scores mean that there are fewer connections tying ideas together, making the text more difficult to process for the reader.
- **Deep cohesion**: This measure reflects the extent to which the text contains causal and intentional connectives when there are causal and logical relationships between concepts in the text. These connectives help the reader form a more coherent and deeper understanding of causal events and processes the text is explaining. Even if a text contains many relationships but lacks these connectives, the reader is required to infer relationships between ideas in the text. This lowers cohesion.
- **LSA overlap**: This measure gives the LSA cosines for adjacent sentences and between all sentences in the text. This measures how conceptually similar sentences are to each other. These LSA measures do not rely on the user to select what comparison text(s) are used. Therefore, these LSA scores will not be the same as those obtained directly from the LSA web service. The three LSA measures used were 1) LSA overlap: adjacent sentences (measures how conceptually similar each sentence is to the next sentence); 2) LSA overlap: all sentences in paragraph (measures how similar each sentence is to every other

sentence); and, 3) LSA overlap: given/new sentences (measures average givenness or newness of each sentence).

- **Connectives**: These elements help in the creation of the cohesive links between ideas and clauses. They also provide clues about text organization. Connectives are such words such as *because, so, and, or, although, first, until, moreover,* and *however,* etc. (Coh-Metrix version 3.0 indices, 2012). We evaluated the incidence of causal connectives and logical connectives.
- **Sentence and word statistics:** In addition to the above Coh-Metrix specific measures, we also compared our hand scores to sentence count, word count, average number of words in each sentence, average number of syllables in each word, and average number of letters in each word.

Results

LSA Analyses

There were significant moderate positive correlations between the LSA score and the overall score given in Study 1 (Preservice: r=0.44, p<0.05; inservice: r=0.38, p<0.05). There were also significant moderate correlations between the LSA Vector cosine value and the overall hand score for preservice teachers (r=0.39, p<0.05) and a strong correlation for inservice teachers (r=0.69, p<0.05). Correlations were lower for the inservice teachers than the preservice teachers for LSA cosines, however the opposite was true for LSA vector lengths, where inservice teachers had the higher correlations; these were all significant group differences. When comparing the groups, LSA cosines had a t-test score of $t(174)=3.92$, p<0.05 and LSA vector lengths had a t-test score of $t(174)=6.70$, p<0.05.

Higher overall hand scores, therefore, were associated with higher cosines. These cosines are indicators of semantic similarity, which verifies that the essays that received higher scores were more semantically similar to the "perfect" essay. These essays were also moderately correlated with the LSA vector lengths, which are indicators of how much information is contained in a text. Higher scoring essays tended to contain more information or explanations that are more thorough. Thus, LSA was able to assess essays based on quality in a similar manner as the human scorers did. This supports the idea that one may use LSA to determine, or at least predict, essay quality (Foltz, 1996; Graesser, Li, & Feng, 2013). However, LSA will not identify mistakes contained in an essay, especially if the mistakes use semantically similar words or words contained nearby in the LSA corpus (i.e., if an essay on cells contains the incorrect idea that "mitochondria manufacture proteins," both "mitochondria" and "proteins" are contained in the corpus of "cells"). An essay could conceivably contain all of the "right" words, but be completely wrong in its facts and LSA would still give a high cosine due to the semantic similarity of co-occurring words. LSA also fails to understand instances of negation or of different meanings of the same word (Kintsch, 2002).

Inservice teachers had higher scored essays, on average, but they were found to be less semantically similar to the model essay than the preservice teachers in terms of the LSA cosine. This is the opposite of what we would expect. However, when considering the LSA vector length correlations with the hand scores, the inservice teachers had a much higher correlation than the preservice teachers. Rehder et al. (1998) found vector lengths reflected general knowledge about the topic and cosines reflect the more narrowly embedded knowledge within the selected comparison texts. This may indicate that inservice teachers used more technical terms and had more information overall in their essays and that preservice teachers used language more similar to the supplied model essay, which was written in response to the same prompt.

Coh-Metrix Analyses

Of all the tested Coh-Metrix measures, the only significant correlations with the overall scores were word and sentence count for each essay (moderately correlated), as well as word length for the preservice teachers. LSA measures contained within Coh-Metrix's output (low to moderate correlations) were also significant for both groups (see Table 3). This indicates that higher scored essays are associated with more words and sentences and had stronger semantic overlap among their sentences. None of the measures of cohesion or connectives was significantly correlated for preservice or inservice teachers (see Table 3).

The lack of significant correlations between the hand scores and the unique Coh-Metrix measures indicates that cohesion measures obtained from Coh-Metrix are not indicative of the quality of essays as determined by human scores. This may be similar to other research that found there was no evidence that higher scored essays are more coherent (McNamara, Crossley, & McCarthy, 2010), which is a major theoretical base for Coh-Metrix. We also ran the correlations for each of the three search tools separately. Because the NSDL map interface visually connects concepts together, there is a possibility that these connections would carry over into the essays and that Coh-Metrix would be able to identify additional cohesion devises. The only significant results were as follows:

Table 3

Correlations between hand-scored overall score and selected Coh-Metrix measures.

Coh-Metrix Measures	PRESERVICE: N=122 Pearson Correlation	Sig.	INSERVICE: N=54 Pearson Correlation	Sig.
Referential cohesion (z-score)	0.05	*p*=0.56	-0.13	*p*=0.34
Referential cohesion (percentile)	0.09	p=0.33	-0.07	*p*=0.62
Deep cohesion (z-score)	0.01	*p*=0.91	0.08	*p*=0.55
Deep cohesion, percentile	0.02	*p*=0.83	0.15	*p*=0.27
LSA overlap: adjacent sentences (mean)	0.22^{*}	*p*<.0.05	0.31^{*}	*p*<0.05
LSA overlap: all sentences in paragraph (mean)	0.23^{*}	*p*<0.05	0.11	*p*=0.43
LSA overlap: given/new sentences (mean)	0.32^{**}	*p*<0.05	0.47^{**}	*p*<0.05
Causal connectives incidence	0.02	*p*=0.85	0.06	*p*=0.68
Logical connectives incidence	-0.06	*p*=0.49	-0.03	*p*=0.85
Sentence count	0.32^{**}	*p*<0.05	0.66^{**}	*p*<0.05
Word count	0.38^{**}	*p*<0.05	0.67^{**}	*p*<0.05
Sentence length, number of words (mean)	0.04	*p*=0.67	-0.14	*p*=0.31
Word length, number of syllables (mean)	0.26^{*}	*p*<0.05	-0.18	*p*=0.20
Word Length, number of letters (mean)	0.28^{**}	*p*<0.05	-0.04	*p*=0.76

* Significant low correlation; ** Significant moderate correlation.

- LSA overlap, Adjacent sentences: Preservice keyword search, $r(41)=0.33$, $p<0.05$
- LSA Overlap, All sentences: Preservice keyword search, $r(41)=0.34$, $p<0.05$
- LSA Given/New sentences: Preservice keyword search, $r(41)=0.38$, $p<0.05$; preservice map search, $r(42)=0.32$, $p<0.05$; inservice Google search, $r(18)=0.58$, $p<0.05$; inservice map search, $r(18)=0.50$, $p<0.05$.
- Sentence count: Preservice Google search, $r(39)=0.35$, $p<0.05$; preservice keyword search, $r(41)=0.33$, $p<0.05$; preservice Map search, $r(42)=0.32$, $p<0.05$; inservice Google search, $r(18)=0.81$, $p<0.05$; inservice maps search, $r(18)=0.77$, $p<0.05$.
- Word Count: Preservice Google search, $r(39)=0.43$, $p<0.05$; preservice keyword search, $r(41)=0.47$, $p<0.05$; inservice Google search, $r(18)=0.81$, $p<0.05$; inservice map search, $r(18)=0.71$, $p<0.05$.
- Word length, number of syllables: Preservice map search, $r(42)=0.31$, $p<0.05$.
- Word length, number of letters: Preservice map search, $r(42)=0.44$, $p<0.05$.

T-tests comparing preservice to inservice teachers show that the only significant differences between the two groups are those dealing with sentence count, word count, word length, and LSA given/new sentences. Inservice teachers had higher correlations for sentence count and word count, and LSA given/new. However, preservice teachers had higher correlations for the measures of word length. There were no significant differences with any of the cohesion measures (see Table 4).

Discussion of Study Two

Latent Semantic Analysis provides an analysis of the semantic similarities between documents and other texts. Coh-Metrix calculates a text's cohesion and the coherent mental representation of that text. We were interested in whether measures generated from these programs were correlated with hand-scored essays. LSA was moderately correlated with the hand scores; however, none of the cohesion measures in Coh-Metrix was correlated with our scores. The correlations between LSA and the hand scores tell us that both can identify essays with different levels of content. Because LSA's cosine is not a measure of correctness but instead is a holistic measure of how

Table 4

T-test results comparing preservice teachers to inservice teachers.

Coh-Metrix Measure	***t*(174)**	**Sig.**
Referential cohesion (z-score)	-0.19	p=0.85
Referential cohesion (percentile)	-0.98	p=0.33
Deep cohesion (z-score)	-0.83	p=0.41
Deep cohesion, percentile	-0.80	p=0.43
LSA overlap: adjacent sentences (mean)	-1.47	p=0.14
LSA overlap: all sentences in paragraph (mean)	-0.72	p=0.47
LSA overlap: given/new sentences (mean)	-2.11	p<0.05
Causal connectives incidence	-1.48	p=0.14
Logical connectives incidence	-0.24	p=0.81
Sentence count	-3.13	p<0.05
Word count	-3.21	p<0.05
Sentence length, number of words (mean)	0.65	p=0.52
Word length, number of syllables (mean)	-4.38	p<0.05
Word Length, number of letters (mean)	-4.36	p<0.05

similar a text is to the sample essay, we could not replace the hand scores for the LSA cosines. In addition, the finding that the correlations for preservice teachers were higher than inservice teachers for the cosines, but lower for LSA vector lengths also needs to be considered. Rehder et al. (1998) suggest that both are indicators of knowledge level, but further research is needed to determine the best way to combine these measures to determine knowledge level of an essay. Our findings support their suggestion for further research in this area.

It would appear from our analysis that Coh-Metrix cannot be used to determine essay quality or level of correctness as we are using it (Crossley & McNamara, 2010, 2011). It is possible that Coh-Metrix picks up aspects of essays that human raters do not. Coh-Metrix is designed to look for the cohesion in a text and this is different from looking for correct information. Part of the rubric designed for this study did ask graders to consider the cohesion of the essay when assigning an overall score. Lower scores were given if there were severe sentence structure problems or if cohesion between ideas was lacking and one of the indicators for high scoring essays was clear cohesion of ideas. Therefore, we would assume that there would be even a minor correlation with Coh-Metrix cohesion measures. Nevertheless, Coh-Metrix was designed to match texts to readers based on how much knowledge they have, and we are asking it to see if we can use the program to see if we can tell the knowledge level of the text writers. It may not be possible to use Coh-Metrix in a way beyond what it was designed to do.

IV. GENERAL DISCUSSION

The purpose of this study was to determine whether the finding of reduced cognitive load and increased efficiency when using NSDL's concept map-based web searches as found by Butcher et al. (2015) would translate into increased learning for preservice and inservice teachers. The implications from linking reduced cognitive load when searching for online content in this manner translates into more learning could have broad implications for teachers and students. The constrained nature of a concept map would help users to both see how concepts are related to each other and to quickly locate relevant material.

In the present study, learning was measured with pre- and postessays. These essays were scored by hand, but given research on the unreliability of human scoring (Attali, Lewis, & Steier, 2012; Breland, Bridgeman, & Fowles, 1999; Foltz, Britt, & Perfetti, 1996; Townsend et al. 1993), we were also interested in whether scores generated by computerized systems were correlated with human scoring. LSA has been found to be as reliable as human scorers in some cases (Foltz, 1996) and Coh-Metrix has been used to show that differences in cohesion levels can led to differences between coherence in readers depending on prior knowledge (Graesser, McNamara, Louwerse, & Cai, 2004).

Overall, the human-based scoring results provide modest evidence participants in our study did learn from their web searches, although this did not differ as a function of search task. They increased the percentage of correct information produced and overall scores from pre-essay to postessay. There was no decrease, though, in the percentage of incorrect statements or in revisions to errors from pre-essay to postessay, which would have been a stronger indication that participants learned from the web searches.

One possible reason for the lackluster results in error revision is the limited amount of time given to participants to both write their answers and to find online resources. Each participant was given five minutes to respond to the initial essay prompt, ten minutes to search for material online, ten minutes to review marked resources to learn from the material, and then five minutes to write the postessay. What we found with many of the essays was that participants were not writing much at all. The average length of the postessays in our study was 138 words for preservice teachers and 161 words for inservice teachers. Most essays with errors only contained a small number of errors (the average number of errors per essay was less than 2). While participants were not making very many errors, they were still not explaining the concepts very thoroughly—on average, preservice teachers had 11 correct idea units in their postessays and inservice teachers had 16. Having more time to construct a response would enable participants to expand on concepts and provide more room for errors in thinking to manifest themselves.

More time searching for content online might also lead to differences in learning gains. One of the goals of the search task was to identify multiple websites that could be used in a classroom setting. It would also help to familiarize users with the map interface, which was unfamiliar to most of our participants. Future studies should examine the differences in learning gains when participants are given more time to both write and search for online content. It appears that participants tended to stick with what they knew or to very broad explanations of the topics instead of branching out into less familiar territory. More time for searching for information, learning from it, and writing the essays might encourage explanations of correct understanding and revisions of errors in thinking to be more thorough.

Our second study evaluated whether overall essay scores are correlated with scores generated by computerized programs. If computerized scores are highly correlated with human scoring, then it is feasible that much of the grading/scoring process could be done automatically. Latent Semantic Analysis, which is designed to determine how semantically similar texts are to each other and to a domain database, was moderately correlated with our hand-scores. This shows that what is determined by human scores to be a more thoroughly explained and correct essay is also assessed as more semantically similar to the "perfect" essay and the domain database contained within LSA. However, these correlations are only moderate (0.44 for the preservice teachers and 0.38 for the inservice teachers). Foltz (1996) found a correlation of 0.68 of LSA to human graders. In this study, however, graders were looking for content overlap between selected texts the essay writers studied and their essays.

Correlations for vector lengths were also significant (0.39 for preservice teachers and 0.69 for inservice). Vector lengths may be a better measure of general knowledge of a topic (Rehder et al., 1998), but both cosines and vector lengths together are indications of semantic similarity. Cosines are measures of the content of the essay; vector lengths refer to the amount of information (Kintsch, 2002). Therefore, at this point, it would not be wise to substitute LSA for the human scorers. Other research has found higher correlations between LSA and human scorers. Foltz (1996) found, after using a weighted mean, a correlation of 0.68 between LSA and expert graders, and, after using multiple settings and manipulations, the highest correlation Pincombe (2004) found was 0.60. However, for both of these studies, the content used in LSA's semantic space was carefully selected based on the topic at hand. For our study, there were not specific LSA semantic spaces (i.e., for "the water cycle") for us to select. Therefore, the semantic space selected was "general-reading-up-to-1st-year-college." This broad category most likely does not contain very many specific resources about the topics at hand and is also filled with much unrelated material. The use of this semantic space may be a contributing factor to lower correlations between LSA and our scores as compared to other studies.

As with Study 1, more time given to participants to allow for longer essays might further determine the compatibility between the two scores. Layfield (2012) found that longer essays have better performance within LSA (i.e., are more reliably scored) than shorter essays, due to the increased amount of information that can be represented in the semantic space. Longer essays may also make automated systems like LSA produce results that are more highly correlated with hand scores. The differences in correlations between the preservice teachers and inservice teachers may simply be a reflection of essay length. Another potential issue with LSA is that it cannot be used to identify errors in essays, particularly when the writer is using the correct vocabulary to explain erroneous ideas.

Of all the Coh-Metrix measures considered, only the number of words and the number of sentences, as well as LSA scores, were correlated with the hand scores. The more specific measures of cohesion and coherence—specifically referential and deep cohesion, as well as connective measures—that are unique to Coh-Metrix were not correlated with our hand scores. This indicates that cohesion measures obtained from

Coh-Metrix are not indicative of the quality of essays as determined by human scores. Coh-Metrix has shown that low-prior-knowledge readers do better with highly cohesive texts to help them form a coherent mental representation; in contrast, high-prior-knowledge readers do better with text that has fewer cohesive cues that allow them to form the mental representation from their prior knowledge (Graesser et al., 2004). Those with very little or no prior knowledge need text with many cohesive cues, because they have little information in their mental representations with which to connect the new information (McNamara, 2001). Our study used groups of participants that could be separated into low and high prior knowledge levels—preservice and inservice teachers. There were no differences in cohesive measures between these groups.

Crossley and McNamara (2010) found that knowledgeable scorers of essays found *less* cohesive texts to be *more* coherent, instead of the other way around. In this study, several coherence measures such as relevance, continuity, and reader orientation were found to be highly correlated with the overall holistic scores of essays. However, these same measures were found to be negatively correlated with Coh-Metrix cohesion measures. (Essays with high levels of coherence were found to have low levels of cohesion). High-knowledge readers are able to fill in the text's gaps with prior knowledge instead of relying on cohesion cues. However, we are interested in the knowledge level of the writers, not the scorers. The lack of either positive or negative correlations between our Coh-Metrix and hand scores shows that neither the presence of cohesion markers (what Coh-Metrix assumes to lead to more coherent texts) nor the lack thereof (which may be an indicator of high knowledge level) is indicative of the essays' scores. It appears that Coh-Metrix is not utilizable for this type of scoring. Again, as previously suggested, longer essays may make a difference. The short nature of our essays may be associated with a general lack of any cohesive cues for both high- and low-scoring essays. Longer essays, where the writer is able to thoroughly explain and explore the essay prompt, would potentially lead to more cohesive cues that may show differences between participants of high and low knowledge.

Butcher et al. (2015) were able to demonstrate that using the map interface was a more efficient search tool in terms of time to evaluate resources, and one with decreased cognitive effort. In the present study, there were only moderate indications that the findings of Butcher et al. translated into increased learning gains. In addition, although LSA cosines were moderately correlated with the overall scores, the results were not strong enough to suggest that LSA could replace human scorers. Coh-Metrix measures were not correlated with the hand scores. Further research should investigate whether more time given to write the essays and to find information online will lead to more robust changes in the essays that will demonstrate that learning has occurred, and whether longer essays lead to higher correlations with computerized scoring systems.

APPENDIX A

NSDL ESSAY SCORING RUBRIC

Each essay will be broken down into idea units, which are determined based on independent clauses or sentences contained within each essay. If some clauses or sentences contain a list of items, then these will be broken down and each one should be evaluated separately.

Each idea unit should be marked as one of the following three things:

1. Correct statements

- A stated, true fact
  - *Volcanoes are found at plate boundaries.*
  - *Lysosome is an organelle.*
  - *The nucleus contains the DNA.*
  - *Heat is needed for evaporation to take place.*
- Repeated concepts are only counted once
  - ***Cells are the basic unit of living things****. They are simple things that make up different creatures and other important organs.*
- Ideas may be paraphrased but still receive full credit

2. Error/Incorrect Statements

- Mark each specific error that appears in the essays. These may be large, misconceptions of the concepts or smaller factual errors
  - *Divergent plate boundaries form mountains.*
  - *The main job of organelles is to take in oxygen.*
  - *Salt water is evaporated.*
- Errors in both the pre- and postessay should be marked and counted.
- Each error in the pre-essay should be evaluated in the postessay according to the error revision instructions.

Error Revision (Post-essay only)

When an error is present in the pre-essay, the postessay is evaluated to see if the error is fixed in the postessay in the following manner:

- **Fixed Error:** The error in the pre-essay has been corrected in the postessay
    - Pre-essay Error: *Salt water is evaporated.*
    - Postessay Fixed Error: *When water is evaporated, salt and other minerals are left behind.*

- **Same Error**: An error from the pre-essay is still present in the postessay.

- **Not-Addressed Error**: An error in the pre-essay is neither fixed nor left present in the postessay. Instead, the matter is completely ignored in the postessay
    - If the pre-essay contains the error about salt water being evaporated, the postessay mentions nothing about what happens when salt water is evaporated.

- Also, mark **New Errors** in the postessay, or errors that are unique to the postessay.

3. Irrelevant/Other Statements

- Superfluous phrases that do not add anything scientific to the essay
- Incomplete ideas/phrases (often at the end of the essay) that do not have enough information to be labeled as correct or incorrect
- Vague ideas or non-scientific terms that are not "technically" incorrect, but neither are they the normally accepted term, (i.e. Calling the nucleus of a cell a “yoke”).
- Off topic phrases: May be correct statements, but are irrelevant to the topic/essay prompt
    - *“As we see in many of the mountains, different layers of earth sediments and minerals have built upon each other over millions of years.“*
    - *If we take a historical prospective, we would be able to use scientific clues to understand more about ancient civilizations using what we know about climate changes effecting water distributions.*

Overall Quality Score

After each idea unit has been scored, evaluate the essay as a whole in the following manner:

*Each essay should be assigned a score between **1 and 5 points.***

1 points: A) No relevant domain content, (i.e. talks about Excel "cells" rather than the biological ones) or only a very small amount of correct information. Essay may also contain many errors and/or serious misconceptions about the topic of interest. Multiple "other" or off-topic statements are often also included (i.e. "I remember studying about plate tectonics in school, but I can't remember much about them.") May make vague references to the topic at hand, but does not explain any concepts with scientific terms.

> I don't know much about cells, as it has been a while since I have taken any science class, but I do know that cells are what make up all things. I think I definitely need to study up before ever trying to teach a unit on cells to students.

***This essay is on topic, but there is no correct information given. Most of essay would be categorized as "irrelevant" or "incorrect."*

2 points: Amount of correct information is minimal (even if all correct). May contain incorrect facts and/or serious misconceptions about topic and will not cover all parts of the prompt. Significant amount of essay may be devoted to off-topic comments. May also demonstrate poor sentence structure/grammar and vocabulary use.

> Water evaporates from bodies of water on earth (lakes, oceans, rivers, etc.) and begins to form clouds. When clouds cannot hold any more water they release it in the form of precipitation.

***This essay contains all correct information, but its short length and lack of any explanation of terms scores it a 2.*

3 points: Mainly relevant domain content, mostly correct statements but are vaguely explained or lacking proper terminology. Low percentage of both "incorrect" and "other" statements may be included. Essay will not cover all aspects of the prompts. It may have problems with sentence structure/grammar (such as poor sentence flow/connections). Cohesion of ideas may lapse at times.

> There are several different kinds of cells. Our bodies are made up of cells. In our bodies these cells are small but have a specific structure. They have a cell membrane that is permeable meaning that things can be transferred in and out of them. They have a nucleus that has the DNA in it that makes up the cell and tells it what kind of cell it needs to be. The cells have parts in them that carries stuff to the nucleus giving it information on what to do. There is a part of the cell that takes care of waste. Its job is to get rid of all the waste that is in the cell. All the organelles of the cell have different shapes as well.

***This essay would receive a 3 because, while most of it is correct, it is lacking correct terms (such as specific organelle names.) The essay reads as choppy, lacks in depth and does not contain a coherent flow.*

4 points: Contains ample domain content that is relatively well explained. May still be missing one aspect asked for in the prompt, but what is included is well explained with only minor error, if any, present. Sentence flow/structure may have some issues.

The crust of the earth is made up of plates that sit on top of the mantle. The mantle is plasma and as a result it is molten and constantly the hotter mass is moving up towards the surface then as it cools it sinks back towards the core. Because of this movement under the solid plates, the plates actually move as much as is possible with the mantle. The plates will hit each other and either crumple at the edges causing upward and downward movement or they will slide past each other. The crumpling movement can create mountain regions and therefore volcanoes. But volcanoes do not have to be tall. They can be ejaculate from cracks that are so low they are allowing the mantle to come right up out of the ground. The sliding movement causes earthquakes and tsunamis. Tsunamis if the earthquake happens below the ocean.

***This essay demonstrates good flow and includes detailed explanations. However, while it mentions different phenomenon that can be caused by the motion of plates, it does not adequately explain the differences between different plate movements.*

5 points: Almost all is relevant domain content that is well explained with proper terms and appropriate connections/relationships. All points of prompt are addressed. Is well organized and demonstrates clear coherence and smooth progression of ideas.

There are two types of tectonic plates on the Earth, land and ocean. Land plates are often made of Granite and are thicker than ocean plates. Ocean plates are often made of basalt and are more dense, sitting lower in the atmosphere and are mainly covered by water. The plates can move in one of three ways; convergent plates move toward one another, divergent plates move away from one another, and transverse plates move alongside one another. When two land or continental plates converge, they often build mountain ranges such as the Himalayas or Alps. When a continental plate converges with an ocean plate, the ocean plate will subduct, or move below, the less dense continental plate. This is true around the Atlantic and Pacific Oceans. When two continental plates diverge, they create a rift valley. When two oceanic plates diverge, they create a ridge such as the Mid-Atlantic ridge where new lava is forced up and spreads the two plates apart. Transverse plate boundaries will often create folded mountains such as the Andes or Rocky Mountains. Volcanoes form when a plate is located over a hot spot, or thin area, of the plate. The lava forces up through the plate and builds upon itself until it forms a volcano or island. Sometimes volcanoes occur when the oceanic plate goes below a continental plate and heat is built up from the melting of the oceanic plate deep in the Earth.

***This essay contains correct terms with adequate amounts of explanation. All parts of prompt are covered and displays good sentence structure and flow.*

APPENDIX B

SAMPLE EDUCATIONAL SEARCH TASK FOR PLATE TECTONICS

TASK 1: PART 1 – 10 minutes

For this task, please imagine that you are **teaching high school science** and focusing on the following Utah Core standard and objective:

Earth Science Standard:

Explain the water cycle in terms of its reservoirs, the movement between reservoirs, and the energy to move water. Evaluate the importance of freshwater to the biosphere.

> **Objective:**
>
> Identify the reservoirs of Earth's water cycle (e.g., ocean, ice caps/glaciers, atmosphere, lakes, rivers, biosphere, groundwater) locally and globally, and graph or chart relative amounts in global reservoirs.

Classroom Information:

This year, you have a number of **students who are lagging behind in science** and **identify themselves as "visual learners."** They don't understand how water changes forms on earth and how this is related to the global reservoirs of the water cycle. You are especially concerned with finding **resources that you can use in small group activities** to help these struggling learners master this standard/objective.

Your Goal:

Select **1-4 digital resources** that you think are **well-matched to the standard and objective** listed above ***and*** **will help your students learn** as they work in small groups during class. You should be sure to choose sites that you think are **high-quality** and **scientifically accurate**.

Directions:

As you search for and evaluate online resources, you'll be making a "yes" or "no" decision about whether you want to save this resource for your students. Please be sure to evaluate the strengths and weaknesses of each resource as you consider it.
If you decide "yes" on a resource, please bookmark it.

Task 1: PART 2 – 10 minutes

Whereas the last task was about *finding* resources, **this task is about *learning* from digital resources**.

Learning Strategy

Now that you have selected one or more digital resources, we'd like you to show us what can be learned from them (even if you already know a lot about the topic). Please explain the materials to yourself by answering the following questions as you explore the digital resources you chose:

- What is the information telling me?
- What does the information mean to me?
- Why is the information important to the topic I am learning?
- How does this information relate to what I already know?
- What questions do I have about the information?

Directions:

If you selected multiple digital resources during the last task, start with the resource you thought was best and work your way down the list – you can move on to the next site at any time you feel ready to go on.

Remember, your goal is to learn as much as you can about:

Earth Science Standard:

Explain the water cycle in terms of its reservoirs, the movement between reservoirs, and the energy to move water. Evaluate the importance of freshwater to the biosphere.

Objective:

Identify the reservoirs of Earth's water cycle (e.g., ocean, ice caps/glaciers, atmosphere, lakes, rivers, biosphere, groundwater) locally and globally, and graph or chart relative amounts in global reservoirs.

APPENDIX C

ESSAY PROMPTS

PLATE TECTONICS: Please write an explanation of what you know about the movement and interactions of plates on the Earth. Please include as much detail as possible about the Earth's plates, how they move, and how plate movements and interactions are related to physical phenomena such as volcanoes and mountains.

CELLS: Please write an explanation of what you know about the structure and function of cells. Please include as much detail as you can about the organelles of a cell, the function of these organelles, and how materials are transported in and out of cells.

WATER CYCLE: Please write an explanation of what you know about the Earth's water cycle. Please include as much detail as you can about the different forms that water can take on the Earth, the global reservoirs of water involved in the water cycle, and how much global water supply is available for human consumption.

APPENDIX D

LSA COMPARISON TEXT FOR THE CELL ESSAYS

Cells make up all living organisms. There are two types of cells: prokaryotic and eukaryotic. Prokaryotes are single-celled organisms that do not contain any organelles. Eukaryote cells comprise animal and plant life and contain organelles. Organelles give structure and allow a cell to perform specific functions. The nucleus may be considered the most important organelle. It is located at the center of the cell and contains the DNA of the cell, which is contained in chromosomes. All cells contain the same DNA, but the type of cell determines what portion of the DNA is active. The nucleolus, a round organelle, is located inside the nucleus and contains the RNA needed for protein manufacture. The nucleus is surrounded by the nuclear membrane. Ribosomes are produced in the nucleus and are sometimes referred to as "miniature protein factories." Ribosomes comprise twenty-five percent of the cell's mass. There are two types of ribosomes; the mobile type floats freely in the cytoplasm, while the stationary type attaches itself to the rough endoplasmic reticulum, another organelle that makes proteins. There is also smooth endoplasmic reticulum, which does not have ribosomes attached and makes lipids (fats). The endoplasmic reticulum (ER) is a transport mechanism within the cell and is connected to the nuclear membrane. Another organelle is the Golgi apparatus, a membrane structure located near the nucleus that packages protein manufactured in the cell. The Golgi apparatus also transports material in and out of the cell. Mitochondria are the "powerhouse" of the cell, creating energy by combining oxygen and sugar to form ATP. ATP is the body's source of energy. Lysosomes break down food for use in the cell and transport waste to the cell membrane for removal. Vacuoles are fluid filled organelles that store water and other materials for the cell. All the organelles are suspended in a jelly-like liquid known as cytoplasm (also called cytosol). This is mostly water but also contains proteins that control the cell's metabolism. The cell membrane (or plasma membrane) holds the cytoplasm and organelles together. It is composed of a lipid bilayer that controls what comes in and out of the cell and is composed of proteins and carbohydrates. Material can enter in one of two ways: passive transport, the process of osmosis and diffusion across the membrane, or by active transport, which involves transport proteins embedded in the cell membrane to let certain substances through.

All of the preceding organelles are found in both animal and plant cells, but there are some organelles unique to each. Only animal cells have centrioles, which are involved in spindle fiber production necessary for cell division. Animal cell membranes are composed of phospholipids, cholesterol, and glycolipids; different animal cells have different ratios of these three. Outside the cell membrane of plant cells is a cell wall, a rigid structure that gives plants their shape. Plant cells also have chloroplasts that create energy for the cell by converting sunlight in to food. They are green due to chlorophyll. Cells produce hydrogen peroxide as a byproduct that is actually poisonous to the cell. Fortunately, they also produce a catalyst to break down the hydrogen peroxide to water and oxygen. White blood cells are able to use the hydrogen peroxide to destroy invading cells in the body.

REFERENCES

Attali, Y., Lewis, W., & Steier, M. (2012). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*(1), 125-141.

Baig, M. (2012). Cohesion Coherence interdependence—Analyzing cohesive devices to study coherence in the text. *Language in India, 12*(10), 98-124.

Braasch, J. L. G. & Goldman, S. R. (2010). The role of prior knowledge in learning from analogies in science texts. *Discourse Processes, 47*, 447-479.

Breland H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework.* New York: College Entrance Examination Board.

Butcher, K. R., Davis, S., & Cook, A. E. (2015, in preparation). Cognitive effects of graphical search interfaces during online information seeking.

Ciullo, S., Falcomata, T. S., Pfannenstiel, K., & Billingsley, G. (2015). Improving learning with science and social studies text using computer-based concept maps for students with disabilities. *Behavior Modification, 39*(1), 117-135.

Coh-Metrix version 3.0 indices (2012). *Institute of Educational Sciences*. Retrieved from http://cohmetrix.memphis.edu/cohmetrixpr/cohmetrix3.html.

Crossley, S. A. & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society,* 984-989.

Crossley, S. A. & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, 1236-1241.

Deniman, D., Sumner, T., Davis, L., Bhushan, S., & Fox, J. (2006). Merging metadata and content-based retrieval. *Journal of Digital Information, 4*(3). Retrieved from http://journals.tdl.org/jodi/index.php/jodi/article/view/105/104.

Dumais, S. T. (2003). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*, 188–230.

Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers, 28*(2), 197-202.

Foltz, P. W., Britt, M. A., & Perfetti, C. A. (1996). Reasoning from multiple texts: An automatic analysis of readers' situation models. In G. W. Cottrell (Ed.) *Proceedings of the 18th Annual Cognitive Science Conference,* (pp. 110-115), Lawrence Erlbaum, NJ.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes, 25*(2-3), 285-307.

Graesser, A. C., Li, H., & Feng, S. (2013). Constructing inferences in naturalistic reading contexts. In E. O'Brien, A. Cook, and R. Lorch, (Eds.), *Inferences during reading*. Cambridge: Cambridge University Press.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. P. Sweet & C. E. Snow (Eds.) *Rethinking reading comprehension,* (pp.82-98). New York: Guilford.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193-202.

Hagemans, M. G., van der Meij, H., & de Jong, T. (2013). The effects of a concept map-based support tool on simulation-based inquiry learning. *Journal of Educational Psychology, 105*(1), 1-24.

Kintsch, W. (2001). Prediction. *Cognitive Science, 25*(2), 173-202.

Kintsch, W. (2002). Latent semantic analysis for machine grading of clinical case summaries. *Journal of Biomedical Informatics, 35*, 3-7.

Landauer, T. K. & Psotka, J. (2000). Simulating test understanding for educational applications with latent Semantic Analysis: Introduction to LSA. *Interactive Learning Environments, 8*(2), 73-86.

Layfield, C. (2012). With LSA size DOES matter. *Computer Modeling and Simulation (EMS), 2012 Sixth UKSim/AMSS European Symposium*, 127-131.

Laiw, S. S. & Huang, H. M. (2003). An investigation of user attitudes toward search engines as an information search tool. *Computers in Human Behavior, 19*, 751-765.

Marchionini, G. & White, R. (2007). Find what you need, understand what you find. *International Journal of Human-Computer Interaction, 23*(3), 205-237.

McIlvain, E. (2010). NSDL as a teacher empower point: Expanding capacity for classroom integration of digital resources. *Knowledge Quest, 39*(2), 54-63.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55*(1), 51-62.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86.

McNamara, D. S., Ozuru, Y., Grasser, A., Louwerse, M. (2006). Validating Coh-Metrix. In R. Sun & N. Miyake (Eds.) *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 573-578). Mahwah, NJ: Erlbaum.

Pincombe, B. (2004). *Comparison of Human and Latent Semantic Analysis (LSA) Judgments of Pairwise Document Similarities for a News Corpus.* Edinburgh South Australia: DSTO Information Sciences Laboratory.

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes, 25*(2&3), 337-354.

Salomon, G., Globerson, T., & Guterman, E. (1989). The computer as a zone of proximal development: Internalizing reading-related metacognitions from a reading partner. *Journal of Educational Psychology, 81*(4), 620-627.

Townsend, M. A. R., Hicks, L., Thompson, J. D. M., Wilton, K. M., Tuck, B. F., & Moore, D. W. (1993). Effects of introductions and conclusions in assessment of student essays. *Journal of Educational Psychology, 85*(4), 670-678.

von der Heidt, T. (2015). Concept maps for assessing change in learning: A study of undergraduate business students in first-year marketing in China. *Assessment & Evaluation in Higher Education, 40*(2), 286-308.

Willoughby, T., Anderson, S. A., Wood, E., Mueller, J., & Ross, C. (2009). Fast searching for information on the Internet to use in a learning context: The impact of domain knowledge. *Computers & Education, 52*, 640-648.